Proc. IAHS, 374, 159–163, 2016
proc-iahs.net/374/159/2016/
doi:10.5194/piahs-374-159-2016

Comparison of cross-validation and bootstrap aggregating for building a seasonal streamflow forecast model

Simon Schick[1,2], **Ole Rössler**[1,2], **and Rolf Weingartner**[1,2]

[1]Institute of Geography, University of Bern, Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, Bern, Switzerland

Correspondence to: Simon Schick (simon.schick@giub.unibe.ch)

Published: 17 October 2016

Abstract. Based on a hindcast experiment for the period 1982–2013 in 66 sub-catchments of the Swiss Rhine, the present study compares two approaches of building a regression model for seasonal streamflow forecasting. The first approach selects a single "best guess" model, which is tested by leave-one-out cross-validation. The second approach implements the idea of bootstrap aggregating, where bootstrap replicates are employed to select several models, and out-of-bag predictions provide model testing. The target value is mean streamflow for durations of 30, 60 and 90 days, starting with the 1st and 16th day of every month. Compared to the best guess model, bootstrap aggregating reduces the mean squared error of the streamflow forecast by seven percent on average. Thus, if resampling is anyway part of the model building procedure, bootstrap aggregating seems to be a useful strategy in statistical seasonal streamflow forecasting. Since the improved accuracy comes at the cost of a less interpretable model, the approach might be best suited for pure prediction tasks, e.g. as in operational applications.

1 Introduction

Small sample sizes challenge the application of statistical models for seasonal streamflow forecasting. For example, a daily hydrometeorological time series of length 30 years can be considered as a long record. However, at seasonal time scales the series provides 30 cases (e.g. summer means). Following the nomenclature described by Hastie et al. (2009), the model building procedure then has to cope with these 30 cases for:

1. model training, i.e. fit models with varying complexity or different predictors;
2. model selection, i.e. validate the models and choose the best one(s); and
3. model testing, i.e. estimate the final models prediction error (possibly by combining several models).

To overcome small sample sizes, resampling is commonly used for model selection and testing. In addition, seasonal streamflow forecasting often encounters weak predictor-predictand relationships, introduced by missing or noisy predictors – e.g. precipitation and temperature of the target season. Models out of any resampling thus can differ markedly, which leads us to the following question: Are there any benefits if the models resulting from resampling are combined in a systematic way? To address this question, we compare (1) the selection of a single "best guess" model along with leave-one-out cross-validation against (2) bootstrap aggregating along with out-of-bag prediction error estimates. Bootstrap aggregating was introduced by Breiman (1996a, "bagging" or "bagged model" for short) and aims to reduce the variance of a statistical model by applying it to bootstrap replicates of the data set and combining the corresponding predictions afterwards.

Below Sect. 2 briefly presents the data set, Sect. 3 outlines the methodology, and in Sects. 4 and 5 the results are presented and discussed, respectively.

Published by Copernicus Publications on behalf of the International Association of Hydrological Sciences.

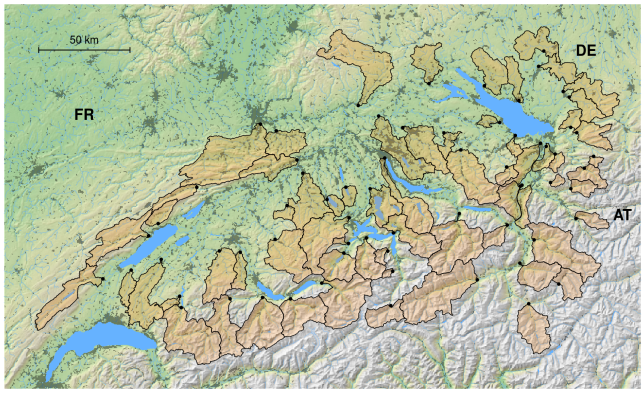

Figure 1. The study region comprises parts of Germany, Austria, and Switzerland. Grey shaded regions indicate urban areas, points mark gauging stations.

2 Data

The hindcast experiment is applied to 66 sub-catchments (no nesting) of the Swiss Rhine at Basel, ranging in mean elevation from 500 to 2300 m and in area from 20 to 900 km^2 (Fig. 1). Streamflow is regulated and routed for the purpose of hydro power, flood protection, water supply, and ecological conservation. Up to 10 catchments can be considered as heavily regulated; for the remaining catchments we assume that anthropogenic effects on the catchments hydrology do not have any impacts at seasonal time scales. Daily mean streamflow in $m^3 s^{-1}$ for the period 1982–2013 is provided by public authorities of Germany, Austria, and Switzerland, whereas daily precipitation and temperature series are catchment averages derived from the E-OBS gridded data set version 12.0 in 0.25° resolution (approximately 19 and 28 km in longitude and latitude; Haylock et al., 2008).

3 Methodology

The comparison of the two model building procedures relies itself on the principle of cross-validation, i.e. some cases are in turn excluded from the data set and the complete model building procedures are conducted by using the remaining cases. Section 3.1 first introduces the regression model, which is common to both model building procedures. Section 3.2 then describes the two resampling approaches; in case of the best guess model, resampling is solely used to estimate the prediction error, whereas in case of the bagged model resampling is at the heart of the model building procedure. Section 3.3 finally states the cross-validation implementation and the statistical test in order to contrast the two procedures.

3.1 Regression model

The regression model follows closely the approach of Garen (1992), i.e. initial conditions are considered only. The predictand $\boldsymbol{y}_{i,j}$ is in turn mean streamflow of duration $i = 30, 60, 90$ d, starting at the 1st and 16th day of every month (date of prediction $j = 1, \ldots, 24$). For a particular choice of i and j, the regression equation is given by

$$\boldsymbol{y}_{i,j} = b_0 \boldsymbol{1} + \mathbf{X}\boldsymbol{b} + \boldsymbol{\varepsilon} \quad (1)$$

where b_0 denotes the intercept, $\boldsymbol{1}$ a vector of ones, $\boldsymbol{b}$ the vector of regression coefficients, and $\boldsymbol{\varepsilon}$ the errors. The $n \times p$ matrix $\mathbf{X}$ has in its $p = 3$ columns antecedent streamflow, antecedent precipitation, and antecedent temperature as predictors; n equals the number of years. The time aggregation is individually selected for each predictor according to Spearman's rank correlation, but has to be one of $10, 20, \ldots, 720$ d. Since these predictors can be highly correlated, the regression coefficients $\boldsymbol{b}$ are estimated using partial least squares (Mevik and Wehrens, 2007). Partial least squares is related to principal components regression, but decomposes the cross-covariance matrix $\mathbf{X}^T \boldsymbol{y}$ instead of the predictors covariance matrix. Regarding model selection, we decide to select at least the first partial least squares direction, as otherwise the regression model shrinks to b_0 in Eq. (1). Please note that we do not make any distributional assumptions about $\boldsymbol{\varepsilon}$.

3.2 Resampling approaches

The regression model from Eq. (1) is applied twofold for a particular catchment and predictand $\boldsymbol{y}_{i,j}$:

1. A single best guess model is selected and the mean squared error of prediction E_{MSP} is estimated according to leave-one-out cross-validation.

2. For each of 100 bootstrap replicates of the data set, one model is selected. These models are then combined by simply averaging their predictions (bootstrap aggregating; Breiman, 1996a). Here, out-of-bag predictions (Breiman, 1996b) are used to estimate E_{MSP}.

Breiman (1996a) showed that the aggregation of unstable models can help to decrease the prediction error. Instability (or high model variance) refers to the case when small changes in the data set lead to large changes in the final estimated model. A simple linear model fitted by ordinary least squares can be considered as an example for a stable model whereas neural networks or regression trees generally are examples for unstable models. The present model from Sect. 3.1 is in our view a stable model – it is linear, consists of three predictors (where only the time aggregations are allowed to vary), and partial least squares further tries to reduce the dimensionality of the predictor space.

The out-of-bag approach is closely related to the leave-one-out procedure in that one case is left out at a time,

i.e. the model averaging considers only those models for which the "left-out" case was not included in the corresponding replicates. Since prediction error estimates of out-of-bag and leave-one-out approximately converge with an increasing number of bootstrap replicates, the out-of-bag estimate provides a convenient alternative – testing a bagged model via leave-one-out can be computationally expensive due to the involved bootstrap. Finally, the choice of 100 replicates is based on the recommendation by Hastie et al. (2009) that model training can be stopped as soon as the out-of-bag error has stabilised.

Hereafter, the two approaches are named BGS/LOO (best guess model BGS in combination with leave-one-out LOO) and BAG/OOB (bagged model BAG in combination with out-of-bag OOB), respectively.

3.3 Hindcast experiment

In order to contrast BGS/LOO with BAG/OOB, the 32 year period of investigation is used for an additional leave-one-out cross-validation. Doing so, we get an estimate of E_{MSP} independently of LOO and OOB. Here, also the three adjacent years of the left out case are omitted to avoid spurious skill due to catchment memory (hence $n = 25$ in Eq. 1). Since BGS/LOO and BAG/OOB are nested inside this "buffered" leave-one-out cross-validation, we refer to the latter as the outer cross-validation. Considering a particular catchment and predictand $\boldsymbol{y}_{i,j}$, three steps are applied:

1. $\boldsymbol{y}_{i,j}$ is centred to mean 0 and scaled to standard deviation 1 with respect to the period 1982–2013.

2. Each year (together with its three adjacent years) is left out once, while the remaining years are used for the application of the model building procedures BGS/LOO and BAG/OOB.

3. The mean value of $\boldsymbol{y}_{i,j}$ serves as a competing model (hereafter named the seasonal regime, SRG). Analogue to BGS, E_{MSP} is estimated by LOO as well as the outer cross-validation.

Paired differences of E_{MSP} are then used for inference. Here, paired differences are calculated such that E_{MSP} of the more complex model is subtracted from E_{MSP} of the less complex model (always per catchment). The mean difference μ is used for a right-sided t test (alternative hypothesis $\mu > 0$). Also a nonparametric bootstrap is applied to estimate the probability $P\{\mu > 0\}$, since the differences not necessarily follow a Gaussian distribution.

4 Results

The results are arranged in three sections: Firstly, we contrast BGS with BAG in order to see whether bagging improves the predictions. Secondly, model skill is evaluated by comparing

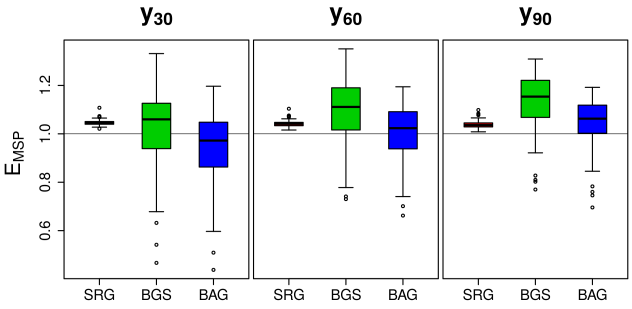

Figure 2. E_{MSP} for all catchments and predictands as obtained from the outer cross-validation; $n = 66$.

BGS and BAG against SRG. Thirdly, we analyse the accuracy of E_{MSP} estimates from LOO and OOB. In the following subscript j is dropped when error statistics are averaged over j.

4.1 Comparison of prediction error

The comparison of BGS against BAG focuses on E_{MSP} from the outer cross-validation: Fig. 2 suggests that BAG scores on average the smaller E_{MSP}. Also the p values indicate that BAG is most likely able to reduce E_{MSP} (third row in Table 1). Table 2 lists additionally E_{MSP} of BGS and BAG for $\boldsymbol{y}_{30}$, $\boldsymbol{y}_{60}$, and $\boldsymbol{y}_{90}$, averaged over all catchments (i.e. the mean value of the corresponding whisker boxes in Fig. 2). Independently of the predictand, reduction of E_{MSP} by using BAG instead of BGS amounts to 7 to 8 %.

4.2 Comparison of model skill

For the evaluation of model skill we focus again on the E_{MSP} estimates from the outer cross-validation (Fig. 2). Due to standardisation of $\boldsymbol{y}_{i,j}$, the benchmark model SRG shows an E_{MSP} near 1 for all catchments (a perfectly estimated mean value would yield an E_{MSP} of 1). On average SRG is a serious competitor and outperforms BGS and BAG in several catchments. Reduction of E_{MSP} by using BGS and BAG instead of SRG is strongest for $\boldsymbol{y}_{30}$ and weakest for $\boldsymbol{y}_{90}$. These findings are also supported by Table 1, which reports the p values of the t test and the bootstrap: It is questionable to unlikely that BGS reduces E_{MSP} on average, whereas BAG very likely does for $\boldsymbol{y}_{30}$ and $\boldsymbol{y}_{60}$, but not for $\boldsymbol{y}_{90}$.

4.3 Comparison of prediction error estimation

Figure 3 shows the differences in E_{MSP}, when LOO and OOB estimates are subtracted from the estimates obtained in the outer cross-validation, which are here considered to be the reference. Thus, a positive difference can be attributed as an underestimation and a negative difference as an overestimation of the prediction error. Apart from a few outliers, the differences lie in the interval $[-0.1, 0.1]$ and are symmetri-

Table 1. p values for the null hypothesis "the simple model outperforms the complex model", estimated by a right-sided t test (paired differences with mean difference μ and alternative hypothesis $\mu > 0$). Paired differences here follow the rule that E_{MSP} of the more complex model is subtracted from E_{MSP} of the less complex model, as specified in the first column. In parentheses also the probabilities $P\{\mu > 0\}$ according to a nonparametric bootstrap with 10 000 replicates are listed. Only E_{MSP} from the outer cross-validation is considered; $n = 66$.

	y_{30}	y_{60}	y_{90}
SRG-BGS	0.1 (0.09)	0.99 (0.99)	0.99 (0.99)
SRG-BAG	<0.01 (<0.01)	<0.01 (<0.01)	0.44 (0.45)
BGS-BAG	<0.01 (<0.01)	<0.01 (<0.01)	<0.01 (<0.01)

Table 2. E_{MSP} of BGS and BAG from the outer cross-validation, averaged over all catchments; E_{MSP} is based on centred and standardised $y_{i,j}$. The last row indicates the reduction in E_{MSP} when BAG is used instead of BGS.

	y_{30}	y_{60}	y_{90}
BGS	1.02	1.09	1.13
BAG	0.95	1.00	1.04
1 – BAG/BGS	0.07	0.08	0.08

cally centred around zero – on average neither LOO nor OOB tend to optimism or pessimism. The heavy negative outliers correspond to the same catchment, which turns out to be regulated due to hydro power.

5 Discussion

In the present study, a hindcast experiment was conducted that mimics the operational use of a simple forecasting system. The objective was the comparison of two model building procedures, which both rely on the same regression model, but use different resampling strategies: A single best guess model, which is tested by leave-one-out cross-validation, and a bagged model, which employs the bootstrap technique in order to build an ensemble of models. An useful byproduct of bagging is the out-of-bag prediction error estimate, which in theory can replace an additional resampling. Regarding the methodology, several points need some attention:

- Catchments were selected without a priori reasoning about their adequacy for seasonal streamflow forecasting. Strictly speaking, none of these catchments exhibits natural streamflow, though some anthropogenic effects might be averaged out due to the seasonal time scale. However, most of these effects are hardly quantifiable and it is not clear whether or not they favour model skill.
- The standardisation of $y_{i,j}$ attaches all seasons and catchments equal weights for the analysis. Doing so, model skill in seasons/catchments with large streamflow variability is masked, e.g. in spring when snow melting occurs and the models perform best (not shown).
- In order to compare the model building procedures, E_{MSP} estimates from the outer cross-validation are considered as the "true" values. This assumption is indeed critical, but unavoidable in the present context – otherwise the real-world data set has to be replaced with a synthetic one.
- The residual analysis (not shown) reveals that the prediction errors are not independent and identically distributed. High flow is commonly underestimated, whereas low flows are often overestimated. Technically, the model can be considered as misspecified, since it lacks relevant predictors (most likely precipitation and temperature during the season to predict). Therefore, common techniques to estimate prediction intervals are not applicable. It remains to be tested whether a substitution of the missing predictors by climate indices or seasonal climate predictions mitigates model misspecification.

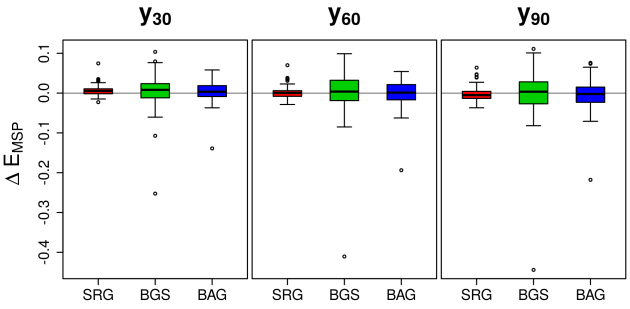

Figure 3. E_{MSP} estimates of LOO (in case of SRG and BGS) and OOB (in case of BAG) subtracted from E_{MSP} estimates of the outer cross-validation for all catchments and predictands; $n = 66$.

6 Conclusions

The results are valid only for the present data set, though the sample size of 66 catchments in combination with 72 predictands might permit more general conclusions:

- BAG scores on average the lower E_{MSP} than BGS. Bagging is useful if the model is unstable (Breiman, 1996a). Since we consider the applied model as rather simple and stable, we argue that instability is introduced by weak predictor-predictand relationships in combination with small sample sizes. These weak (and sometimes spurious) relationships propagate through the screening of the time aggregation, the selection of partial least squares directions, and the final regression coefficients. Small changes in the data set thus often cause that completely different models are identified as the correct one.

- For 30 and 60 day mean streamflow, BAG outperforms in the majority of catchments a naive forecasting strategy, which relies on long-term averages only (SRG). Otherwise it is either questionable (30 day mean streamflow in case of BGS and 90 day mean streamflow in case of BAG) or very unlikely that BGS and BAG provide on average a smaller E_{MSP} than SRG.

- LOO and OOB estimates of E_{MSP} are for most catchments close to E_{MSP} from the outer cross-validation. Neither LOO nor OOB tend to optimistic or pessimistic estimates. Thus, instead of testing the bagged model via the outer cross-validation, also the OOB estimates had been quite accurate.

In practice, statistical seasonal streamflow forecasting is commonly confronted with small sample sizes and weak predictor-predictand relationships due to missing or noisy predictors. The results of the present study indicate that bagging is also able to reduce a pseudo model variance, introduced by weak relationships and intensified by small sample sizes. If resampling is anyway part of the model building procedure and weak relationships come along with small sample sizes, we propose to prefer bagging to the best guess model approach – the computational costs are nearly the same, out-of-bag predictions provide model testing, and prediction errors are likely to decrease. This benefit however comes at the cost of a hardly interpretable model. We thus argue that bagging is most useful when prediction alone is the goal, i.e. in operational forecasting, be it seasonal streamflow or another environmental variable.

7 Data availability

The streamflow series are provided by federal offices and were manually compiled. The corresponding data policies do not allow data dissemination, though for Bayern (http://www.gkd.bayern.de/fluesse/abfluss/karten/index.php?thema=gkd&rubrik=fluesse&produkt=abfluss&gknr=0, GKDB, 2016) and Austria (http://ehyd.gv.at/, BMLFUW, 2016) the series can be accessed online. The E-OBS data set is publicly available at http://www.ecad.eu/ (E-OBS, 2016), the Corine Land Cover at http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-2 (CORINE, 2016), and the EU-DEM at http://www.eea.europa.eu/data-and-maps/data/eu-dem (EU-DEM, 2016).

Acknowledgements. Runoff series and catchment boundaries are provided by the following authorities: Landesanstalt für Umwelt, Messungen und Naturschutz Baden-Württemberg; Bayerisches Landesamt für Umwelt; Land Vorarlberg (data.vorarlberg.gv.at); Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft Österreich; and Schweizerisches Bundesamt für Umwelt. We also acknowledge the E-OBS data set from the EU-FP6 project ENSEMBLES (ensembles-eu.metoffice.com) and the data providers in the ECA&D project (www.ecad.eu). Figure 1 is produced by using Copernicus data and information funded by the European Union (EU-DEM layers) as well as the Corine Land Cover 2006 raster data (version 17) of the European Environment Agency (EEA). Bernhard Wehren made available additional streamflow data for the river Kander at Hondrich. We also thank David M. Hannah for the careful review. The study was funded by the Group of Hydrology, which is part of the Institute of Geography at the University of Bern, Switzerland.

References

BMLFUW: Streamflow monitoring Austria, Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, http://ehyd.gv.at/, last access: 11 August 2016.

Breiman, L.: Bagging Predictors, Mach. Learn., 24, 123–140, doi:10.1023/A:1018054314350, 1996a.

Breiman, L.: Out-of-bag estimation, https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf (10 August 2016), 1996b.

CORINE: Corine Land Cover 2006 raster data, European Environment Agency, http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3, last access: 11 August 2016.

E-OBS: Daily temperature and precipitation fields in Europe, ECA&D – European Climate Assessment & Dataset, http://www.ecad.eu/download/ensembles/ensembles.php, last access: 11 August 2016.

EU-DEM: Digital Elevation Model over Europe, European Environment Agency, http://www.eea.europa.eu/data-and-maps/data/eu-dem, last access: 11 August 2016.

Garen, D. C.: Improved techniques in regression-based streamflow volume forecasting, J. Water Res. Pl.-ASCE, 118, 654–670, doi:10.1061/(ASCE)0733-9496(1992)118:6(654), 1992.

GKDB: Streamflow monitoring Bayern, Gewässerkundlicher Dienst Bayern, http://www.gkd.bayern.de/fluesse/abfluss/karten/index.php?thema=gkd&rubrik=fluesse&produkt=abfluss&gknr=0, last access: 11 August 2016.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer New York Inc., 2. edn., doi:10.1007/978-0-387-84858-7, 2009.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res.-Atmos., 113, d20119, doi:10.1029/2008JD010201, 2008.

Mevik, B.-H. and Wehrens, R.: The pls Package: Principal Component and Partial Least Squares Regression in R, 18, 1–23, doi:10.18637/jss.v018.i02, 2007.